In the name of Allah, the Most Gracious, the Most Merciful









MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
FACULTE DE MEDECINE D'ALGER
DEPARTEMENT DE MEDECINE

STATISTIQUE DESCRIPTIVE

Cours destiné aux étudiants de 6ème année en médecines
Module d'épidémiologie et de médecine préventive

Dr G. BRAHIMI
MAITRE ASSISTANTE EN EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

STATISTIQUE DESCRIPTIVE

Objectif général

- Fournir à l'étudiant des notions de base en statistique descriptive pour le recueil, le traitement et l'interprétation des données recueillies en vue d'aider à la prise de décision dans le domaine de la santé.
- *Apprécier l'importance de vérifier les méthodes de recueil et l'utilisation pratique de celles –ci.*

Objectifs spécifiques :

L'étudiant doit être capable de :

- Définir le type de variables
- Mettre en ordre les données d'un caractère qualitatif et quantitatif
- Traiter les données
- Présenter les données sous forme de tableaux
- Déterminer et apprécier l'intérêt des fréquences cumulées
- Définir, déterminer, calculer les paramètres de tendance centrale, moyenne, médiane, mode
- Définir, calculer les paramètres de dispersion variance, écart type.
- Apprécier les principaux avantages et inconvénients des paramètres de réduction
- Représenter graphiquement les données d'une variable qualitative
- Représenter graphiquement les données d'une variable quantitative
- Représenter graphiquement une distribution à 2 variables qualitatives.
- Calculer, interpréter

Effectifs cumulés, fréquences relatives, fréquences relatives cumulées, moyenne, médiane, mode, variance, écart type, coefficient de variation.

Cours de Statistique descriptive

I / Objectifs et compétences à acquérir
II / Contenu de l'enseignement
III / Supports des cours
IV / Activités d'apprentissage et déroulement des activités et planning
V / Volume horaire :
VI/ Documents et enquêtes mise à la disposition des étudiants sur CD
VII/ Critères d'évaluation :

I / Objectifs et compétences à acquérir

Comprendre le rôle des statistiques dans la prise de la décision médicale1 – VARIABILITE EN MEDECINE
Intérêt des statistiques en Médecine

Intérêt respectif de la moyenne et de la médiane et de l'écart type et du coefficient de variation
Apprendre à : Décrire un phénomène de santé
Renseigner sur des faits pour prendre des décisions
A collecter des données
à les ordonner
à les contrôler
à les traiter
à identifier les différents biais
à représenter les résultats sous forme de tableaux et de représentation graphique et à interpréter les résultats

Apprendre à calculer et à interpréter: pour les variables quantitatives discrètes et continues
les effectifs cumulés
les fréquences relatives
les fréquences relatives cumulées
la moyenne
la médiane
le mode
la variance et écart-type
le coefficient de variation

Plan du cours

1/ Introduction

1.1 Définition : G. Morlat définit dans l'encyclopédie Universalis, la statistique : "Le mot «statistique » désigne à la fois un ensemble de données, d'observation et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation."
Georges Morlat ajoute à ce sujet : " En réalité, ces deux sens du mot statistique ont naturellement entre eux des liens étroits. Il serait vain de recueillir des données, si ce n'était pour les traiter et les interpréter en vue d'éclairer les actions humaines ou de faire progresser la connaissance des phénomènes. Inversement, la manière de recueillir des données peut et doit être influencée d'abord par les méthodes de traitement ultérieures et ensuite par l'utilisation pratique de ces données ou des produits qui en sont dérivés.
On distingue deux domaines :

- **la statistique descriptive**(statistique au sens commun du terme) appelée communément les statistiques.

Elle permet de recueillir, de mettre en forme les données sous forme de tableaux ou de graphiques, de synthétiser l'information recueillie, de la traiter et de l'interpréter afin d'en faciliter la connaissance.

- **la statistique inférentielle.**

Elle permet, à partir d'un nombre réduit d'observations, d'étendre, de généraliser, sous certaines conditions, les conclusions obtenues à la population, d'estimer et de prédire.
La statistique permet de «dénombrer, calculer, mesurer, estimer, juger», et ainsi d'évaluer les phénomènes dans le domaine de la santé.

La statistique aide dans 3 domaines
– Comment recueillir les données ?
– Comment résumer et analyser ces données?
– Avec quelle précision ?

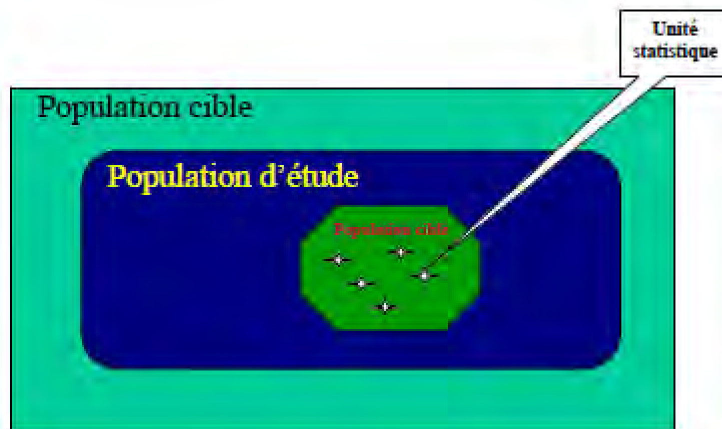
1.2 Séquences d'un projet

Planification
Plan d'analyse Comment recueillir les données ?
Recueil des données

Data management
Analyse des données
Présentation des résultats Comment analyser et résumer les données?
Interprétation
Publication des résultats Avec quelle précision?

1.3 Principaux problèmes qui se posent dans la préparation de l'enquête:

– la définition de la population cible, source, d'étude
– la définition de l'unité de base (unité statistique)
– la définition des observations à réaliser
– le choix d'une méthode d'échantillonnage
– la détermination de la taille de l'échantillon
– le choix de la méthode de collecte des données



2/ Recueil des données - sources d'information

2.1 Le recensement : Le recensementconsiste à recueillir les données démographiques, économiques, et sociales se rapportant à un moment donné à tous les habitants d'un pays donné.

2.2 Le sondage : Le sondage consiste à déterminer certains caractères d'une population à partir des résultats obtenus sur une partie de la population, appelée échantillon

La définition d'une procédure d'échantillonnage repose sur les principes suivants
La définition de l'unité de sondage
La taille de l'échantillon
Le choix des individus

Les organisations professionnelles liées aux entreprises
Les organismes nationaux

Les différents ministères qui publient régulièrement, des études, des rapports statistiques :le Ministère de la Santé, de la Population et de la réforme hospitalière (MSPRH)au niveau central, l'institut national de statistique,le centre national de documentation, la caisse nationale de sécurité sociale,l'Institut National de Santé Publique (INSP), l'Institut Pasteur (IP), l' Agence Nationale de documentation en santé(ANDS), le Centre national de toxicologie, l'Agence Nationale du Sang, l' Ecole Nationale de Santé Publique, l'Institut national de pédagogie, le Laboratoire National

de contrôle des produits pharmaceutiques, le Centre national de pharmacovigilance (Formation à la prescription rationnelle des médicaments)

Les organismes internationaux : L'Organisation des Nations Unis (ONU), l'Organisation Mondiale de la Santé (OMS),l'Organisation des Nations Unis pour l'Education, la Science etla Culture (UNESCO), le Bureau International du Travail (BIT),laCommunauté Economique Européenne(CEE)

Les organismes para officiels et privés

Les enquêtes et questionnaires

Lorsque les sources d'information sont insuffisantes, on peut envisager d'entreprendre des études particulières compte tenu des objectifs retenus.

2.3 Biais

Pour tirer de conclusions valables d'étude statistique il faut identifier les biais. Ils peuvent être liés à des :

- Erreurs d'échantillonnage (intervalle de confiance, degré de précision recherché. Plus la précision est élevée, plus la taille de l'échantillon doit être étendue.Le degré de confiance accepté généralement est de 95% c'est-à-dire qu'on accepte qu'il y ait 5 chances sur 100 pour que la vérité s'écarte de plus de x% du résultat trouvé.
- Erreurs dans le recueil des données

Erreurs dues aux questionnaires (questionsmal rédigées)

Erreurs dues aux enquêtés (oubli, crainte d'avouer son ignorance, ne pas communiquer toute l'information dont il dispose

- Erreurs dues aux enquêteurs (mauvaise maitrise du questionnaire)

<u>3/ Notions préliminaires - Définitions</u> :

3.1 Effectif

L'effectif total est le nombre d'individus dans la série.

L'effectif ou fréquence absolue est le nombre d'individus appartenant à une modalité donnée.

Prenons l'exemple de la distribution de 50 malades selon le sexe. Parmi ces 50 malades, 15 sont de sexe masculin et 35 de sexe féminin. Les effectifs correspondant à chacune des deux modalités sont 15 et 35.

3.2 Fréquence relative

La fréquence relative est le rapport entre l'effectif de la classe ou de la modalité et l'effectif total de la série, le numérateur fait obligatoirement partie du dénominateur.

Elle est exprimée le plus souvent en %.

Pour l'exemple précédent, la fréquence relative du sexe masculin est :

15/50 = 0.30 = 30 % et la fréquence relative du sexe féminin est : 35/50 = 0.70 = 70 %

3.3 Ratio

Un ratio est le rapport des fréquences (effectifs ou fréquences relatives) de deux modalités d'une même variable. Le numérateur n'est pas compris dans le dénominateur. Généralement le numérateur et le dénominateur se réfèrent à deux catégories d'une même variable.

Ex : sex-ratio = effectif masculin
effectif féminin

3.4 Taux

Le taux mesure la probabilité de survenue d'un évènement donné au cours du temps. Un taux doit toujours s'exprimer en fonction d'une certaine unité de temps, pour un lieu géographique donné et pour un groupe de personnes bien défini.

Le numérateur est un nombre d'évènements (décès, maladie, handicap) survenus au cours d'une certaine période t1-t2. Le dénominateur représente la population exposée au risque de survenue de cet évènement pendant cette période.

Ex : si dans les 24 h qui suivent un repas à une cantine fréquentée par 300 personnes, 30 présentent des signes d'intoxication alimentaire, le taux de la maladie est : 30/300 = 0.10 =10.0 % = 100 pour mille.

3.5 Indice

Un indice est une pseudo - fréquence relative ; c'est un substitut d'une fréquence relative difficile à calculer. (Le numérateur n'est pas compris dans le dénominateur).

L'indice est utilisé quand le dénominateur est difficile à déterminer.

3.6 Modalité

Une modalité est une catégorie de la variable étudiée ou une valeur qui affecte un caractère

Exemple :

Le sexe est une variable à deux modalités masculin et féminin, la maladie est une variable parce qu'on peut lui définir au moins deux modalités : malade et non malade, la glycémie est aussi une variable puisque ce caractère présente différentes valeurs.

Il est primordial de bien définir les modalités des caractères de l'étude pour éviter les difficultés qui ne manqueront pas de surgir lors de la collecte et de l'exploitation.

3.7 Variable

La variable est une caractéristique des individus constituant la série étudiée, c'est un facteur susceptible de prendre une valeur différente selon les individus étudiés.

Elle est mesurée ou étudiée sur chacun des éléments de l'échantillon.

C'est une entité qui peut prendre toutes les valeurs d'un ensemble appelé domaine de la variable.

La variable peut être :
aléatoire, dans ce cas la valeur que peut prendre la variable est soumise aux lois du hasard
déterministe : la valeur que peut prendre la variable n'est pas soumise aux lois du hasard.

On répartit les variables en deux (02) groupes principaux :
- les variables qualitatives,
- les variables quantitatives.

Variables quantitatives : mesurables sur une échelle… avec une unité
• de valeurs réelles : **donnée continue :** poids, taille, âge, PA, glycémie, ...
ou
• de valeurs entière : **donnée discrète :** nombres d'enfants, de métastases, nombre de lits etc ...
Les variables quantitatives peuvent être transformées en variables ordinales puis en variables qualitatives ; une variable qualitative ne peut pas être transformée en variable quantitative : poids (kilogramme - ou poids léger, moyen, lourd..).

Variables qualitatives : non mesurables sur une échelle (notion dejugement), mais ...
S'expriment par des mots

Variables catégoriques (nominales)
– homme/femme
– marié/célibataire
– fumeur/non-fumeur
– localisations tumorales
– groupes sanguins : A/B/AB/O

Parfois avec relation d'ordre :donnéesordinales (semi-quantitative)
Exemple : intensité d'une douleur (0, +, ++, +++)
Stades d'un cancer : I, II, III, IV
Consommation de tabac : absence, modéré,important

• Attention : une donnée ordinale n'est pas unedonnée quantitative discrète
– ex :– Stade Cancer du sein : I, II, III, IV
– Rang des enfants : 1, 2, 3, 4, 5+
– Intensité d'une douleur (0, +, ++, +++)

Échelles de classification
Echelle nominale :
Les classes sont mutuellement exclusives sans ordre déterminé elles sont

dichotomiques. (Exemple : Infecté/non infecté ; rural /urbain)
Echelle ordinale :
Les classes sont mutuellement exclusives sont nommées mais dans un ordre déterminé qui correspond à l'ordre naturel existant entre les classes. Exemple : (niveau de revenu, niveau de scolarité.)
Échelle numérique :
Les classes sont représentées par des valeurs numériques ou des valeurs mesurées. Exemple : âge, taille

3.8 Population (encore appelée ensemble)

C'est l'ensemble de tous les individus dont on cherche à déterminer une ou plusieurs caractéristiques, chaque individu est distinct des autres.
L'individu (encore appelée élément ou unité statistique) est l'entité élémentaire de base observée par le statisticien. Cela peut être une personne, un animal, un événement.

3.9 Série (ou distribution) :

C'est un ensemble de données (de valeurs) relatives à une variable mesurée sur un échantillon ou une population d'éléments. La valeur est la mesure.

3.10 Echantillon

• Échantillonnage aléatoire : tirage au sort de façonindépendante, des unités statistiques del'échantillon. Chaque élément de la populationdoit avoir la même probabilité d'être tiré au sort
•Échantillon représentatif **:** Échantillon qui doitrefléter la composition de la population.
• 4 – On mesure la variable : On obtient un ensemble devaleurs, qu'on appelle Série Statistique
– **EXEMPLE** : Sur ce tableau chaque point représente une valeur de la glycémie



• Toutes les valeurs de cette série mesurent une VARIABLE (ici la glycémie)

Mesures biomédicales

• permettent d'étudier des phénomènesbiologiques
– distinguer le "normal" du "pathologique"
– mesurer l'évolution d'une maladie

Exemple : TA, Temps, Glycémie, Poids, Taille, Douleur etc…………… ce sont des variables !

4/ Démarche statistique

• 1- On cherche à caractériser un phénomène qui concerne une population donnée:
– **Exemple : Prévalence du diabète**

• 2- On résume ce phénomène à la mesure d'une ou plusieurs variables mesurée sur 1 unité statistique

- Données qualitatives (identification, âge, sexe, profession, pathologie principale et associées etc)
- Données quantitatives (glycémie, Hémoglobine, glycémie etc …)

• **3 -** Puisqu'il est difficile de faire les mesuresexhaustives sur l'ensemble de la populationconcernée, on se restreint à un sous ensemble(échantillon)

Série statistique : ensemble des mesuresrecueillies pour une variable donnée
Que va-t-on faire de cette série statistique ?



–Analyse descriptive : Résumer et présenter les donnéesrecueillies afin de faciliter leur lecture :tableaux et graphiques

–Analyse inductive ou inférence•généraliser les conclusionsobtenues sur l'échantillon à lapopulation cible en acceptantun certain risque de se tromper

• On doit être en mesure de déterminer ces 5paramètres pour toute étude statistique
• Population cible : Pop Algérienne
• Population d'étude Pop diabétique hospitalisée
• Echantillon 3400 personnes hospitalisées
• Variable : Glycémie à jeun
• Série statistique les valeurs obtenues par les mesure

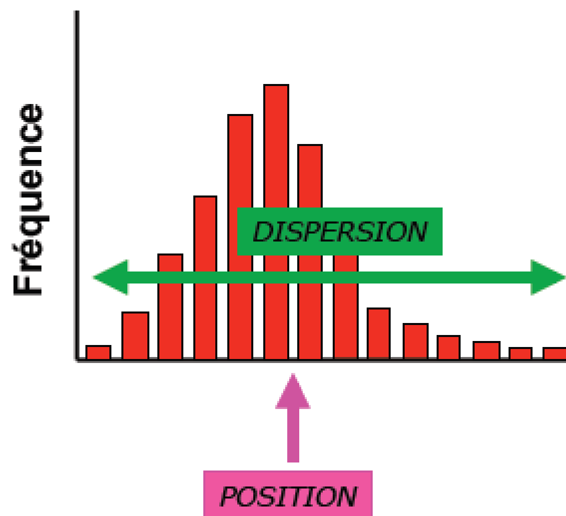
4.1 Présentation des données:

Statistique descriptive : Pour décrire les données, onpeut

- Établir des tableaux
- Regrouper les données dans des classes
- Dessiner des diagrammes

Pour résumer les données afin de les exprimer ou les comparer

- On calcule des paramètres (ou indicateurs)
  - ✓ **De POSITION**
  - ✓ **De DISPERSION**



Paramètres de position

– Médiane
– Mode
– Moyenne
– Quartiles, déciles, percentiles

Paramètres de Dispersion

– Amplitude = Entendue (Minimum, Maximum)
– Intervalle interquartile
– Variance
– Écart type
– Coefficient de variation

Caractéristiques de tendance centrale

Il s'agit de définir une valeur c autour de laquelle serépartissent les observations
Les plus utilisées sont la médiane, la moyennearithmétique et le mode

Notation sur les distributions

• Une distribution est donnée par une série decouples (indice (valeur de la variable) et l'effectif de cetindice)
• Contrairement aux séries statistiques, il estimplicite que les valeurs soient rangées parordre croissant.

Le classement des données : une nécessité

Indice (valeur de la variable)	Effectif
4,9	18
5,0	322
5,1	211
5,2	189
.	.
.	.

Un tableau montrant les couples (critère,fréquences) s'appelle un tableau de distributionde fréquences

Classes	F
0-4	5
5-9	5
10-15	6
16	1
17	1
18-19	2
20-24	5
25-59	35
60-79	20

Tableaux de distribution de fréquences : Il est plus commode de regrouper les données enquelques classes.

Exemple:
Glycémie **Nombre de personnes**
Moins de 5 mmol18
5 à 5,9 mmol 2616
6 à 6,9 mmol426
7 à 8 mmol241
sup à 8 mmol 99

4.2 Regroupement des valeurs en classes

Pour présenter les données d'une variable quantitative sur un nombre important d'individus, on procède à la transformation d'une variable quantitative continue en une variable quantitative discrète (classes) ou d'une variable quantitative discrète en une variable qualitative ordinale (ordre) ; cette opération fait perdre de l'information.
Pour pouvoir dresser un tableau d'effectifs qui soit facile à lire, il faut grouper les observations ou données dans un certain nombre de classes successives, contigües ne se recouvrant pas.
Chaque classe est définie par ses limites, son amplitude et sa valeur centrale.
Les limites de classes doivent être bien précisées. On définit des bornes (limites) de classe, la borne inférieure est la plus petite valeur, la borne supérieure est la plus grande, chaque classe comprend toutes les valeurs égales ou supérieures à sa borne inférieure mais uniquement les valeurs inférieures à sa limite supérieure.
Une observation ne doit être située que dans une seule classe.
C'est pour cela que la convention adoptée est de toujours inclure dans la classe la limite inférieure et donc de toujours exclure la limite supérieure.
Les classes sont contiguës, sans chevauchement, borne inférieure comprise, borne supérieure exclue.
L'amplitude de classe est la différence entre la borne supérieure et inférieure de la classe,

Le nombre de classes est en général déterminé de façon arbitraire ou en utilisant des règles. Il n'existe pas de règle permettant d'imposer le nombre de classes.
Souvent, on s'efforce de construire des classes d'amplitude égale. Si on choisit une amplitude égale pour les classes, la règle suivante a pu être proposée : Nombre de classes = étendue / amplitude.
L'étendue (ou marge) est la différence entre la valeur la plus grande et la valeur la plus faible de la série.

Comment déterminer le nombre de classes
Dans le cas des variables continues,
le choix des intervalles de classe est délicat:
- Trop petits: le nombre de classe est trop grand pourêtre maniable
- Trop grands: des détails sont dissimulés au sein d'unemême classe

Variables continues : La plupart des études sont réalisées avec :
Des intervalles de classes d'amplitude aussi égales que possible
Les classes de fréquence nulle sont évitées
Calculer le nombre de classes
La règle de STURGE : Nombre de classes = 1+3logN
La règle de YULE : Nombre de classes = $2,5\ N^{1/4}$
En général on utilise K= racine carrée de N
Nombre de classes = Etendue/amplitude (même amplitude)

Définir les bornes, les intervalleset les indices des classes
•Borne inférieure : la plus petite valeur de la classe
• Borne supérieure : la plus grande valeur de la classe
• Intervalle de classe : Borne sup – Borne inf
• Indice de classe : valeur centrale de la classe
Exemple : on a relevé le poids de 19 étudiants. L'unité de poids retenue est le kilogramme et les résultats sont les suivants :

76,340	60,400	68,280	57,740	64,990	83,450	79,650	64,100
72,880	69,120	59,990	61,820	61,820	76,360	66,330	52,990
70,560	70,130	65,450					

L'étendue dans l'exemple du poids est : 83.450 – 52.990 = 30.460 kg
Si on décide de prendre une amplitude de 5kg, le nombre de classes est environ de : 30.46 / 5 = 6,09 classes.
Souvent, on s'efforce de construire des classes d'amplitude égale

Le tableau d'effectifs peut être présenté comme suit lorsqu'on retient 7 classes

Poids (Kg)	Centre de classes	Effectif	%
50-54	52.5	1	5.3
55-59	57.5	2	10.5
60-64	62.5	5	26.3
65-69	67.5	4	21.2
70-74	72.5	3	15.8
75-79	77.5	3	15.8
80-84	82.5	1	5.3
Total		19	100.0

4.3 Présentation des données en fonction du type de variables

• Variables qualitatives : Distributions de fréquence
• Variables continues : Mesures de position/centralité
– Mesure de la variabilité/dispersion

4.3.1 Présentation des données Variable qualitative : Distributions de fréquence
• Tableau (ou distribution de fréquence) quidonne le nombre (le %) d'individus selon lesvaleurs de la mesure
• Moyen le plus simple de caractériser lesvariables qualitatives
• Fréquences relatives ou cumulées
Les différentes fréquences
• Fréquence absolue
• Fréquence relative
• Pourcentage
• Fréquence cumulée d'une classe
• Fréquence relative cumulée
• Pourcentage cumulé

Distribution des décès selon le diagnostic (CH …..1987)

Diag	Freq absolue	Freq cumulée	Freq relative	Pourcent	Freq rela cumulée	% cumulée
Choc sept	6	6	0,100	10,0	0,100	10,0
Infarctus	6	12	0,100	10,0	0,200	20,0
CIVD	21	33	0,350	35,0	0,550	55,0
Pneumo	9	42	0,15	15,0	0,700	70,0
Mal inf	10	52	0,167	16,7	0,867	86,7
Lymph	6	58	0,100	10,0	0,967	96,7
Autres	2	60	0,033	3,3	1,000	100,0
total	60	60	1,000	100,0	1,000	100,0

4.3.1.1 Représentation tabulaire

• Un tableau doit contenir toutes les indicationsutiles à sa compréhension
• Un tableau doit contenir: - Un titre (quoi, qui, quand, sources desdonnées ?)
- Les données (présentation)
- Le total des effectifs de la série et le total des fréquences
- Les unités de mesure doivent être mentionnées en tête
- Les codes, abréviations, symboles sont expliqués en bas de la ligne ou de la colonne correspondante
- Les codes,abréviations,symboles sont expliqués en bas de page de même que la source pour des données originales

• quelques termes :

Effectif (= fréquence absolue) = nombre d’observationdans une classe

Fréquence relative =effectif classe i / effectif Total

Fréquence Cumulée = somme de la fréquence d’uneclasse et de toutes celles qui la précède

Fréquence Cumulée relative = somme de la fréquence d’uneclasse et de toutes celles qui la précède / effectif Total

Tableau statistique à simple entrée : Cas d'hépatite C enregistrés dans les Daïras de la wilaya de Souk Ahras en 2015

Communes	Fréquence	Fréquence relative (%)	Fréquence cumulée	FréquenceCumulée relative (%)
Bir Bou Haouch	80	25.4	80	25.4
Heddada	15	4.8	95	30.2
M'daourouch	31	9.8	126	40
Mechroha	33	10.4	159	50.4
Merahna	15	4.7	174	55.1
Ouled Driss	11	3.5	185	58.6
Oum El Adhaim	25	7.9	210	66.5
Sedrata	26	8.2	236	74.7
Souk Ahras	39	12.4	275	87.1
Taoura	40	12.7	315	99.8
TOTAL	315	100	315	100

La wilaya de Souk Ahras compte 10 daïras.

Tableau statistique à simple entrée

Tranches d'âge (ans) / Classes poids (kg)	3-4	4-5	5-6	Total
< 15	19	7	1	27
15-20	32	21	12	65
20-25	3	18	28	49
25-30			1	1
Total	54	46	42	142

Que manque-t-il à ces tableaux ?

Prévalence des infectés et des infections, par type d'établissement

Type d'établissement	Patients	Infectés		Infections	
	N	N	%	N	%
CHU	8505	402	4.73	447	5.26
EPH	1347	90	6.68	93	6.90
EHS	726	13	1.79	13	1.79
Total	10578	505	4.77	553	5.23

Diagnostics d'entrée en réanimation

Symptôme N (%)	Décédés N = 10	Survivants N = 59	Total N = 69
Coma	3 (30.0)	22 (37.3)	25 (36.2)
Choc	6 (60.0)	8 (13.6)	14 (20.3)
Convulsion	0	12 (20.3)	12 (17.4)
Acidose	1 (10.0)	9 (15.3)	10 (14.5)
IRA	0	5 (8.5)	5 (7.2)
CIVD	0	3 (5.1)	3 (4.3)

4.3.1.2 Représentation graphique : variablequalitatives

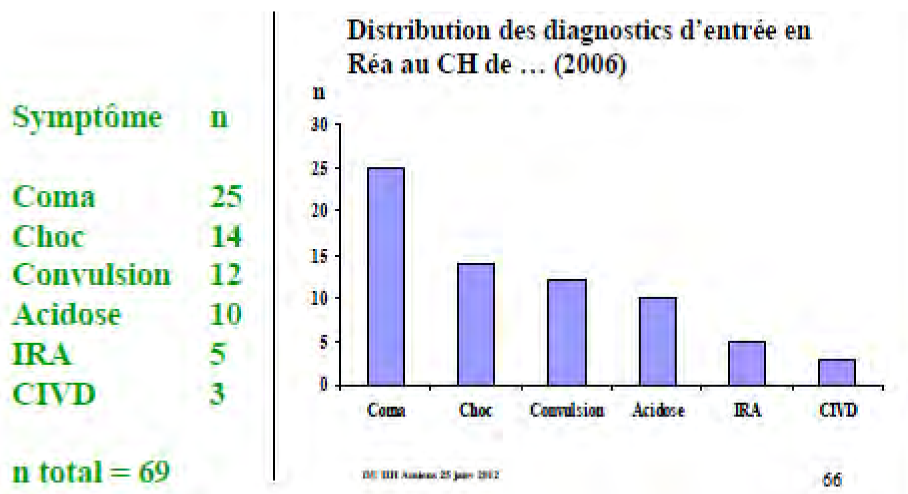
Variables quantitatives comptées et qualitatives :de préférence Diagrammes

DIAGRAMMES:pour les variables quantitatives comptées et les variables qualitatives

• Le plus répandu : diagramme enbâtons (ou en barres)

• pie (camembert, tarte, …) : quand onfait référence au TOTAL de 100%

Distribution de fréquences Diagramme en bâtons



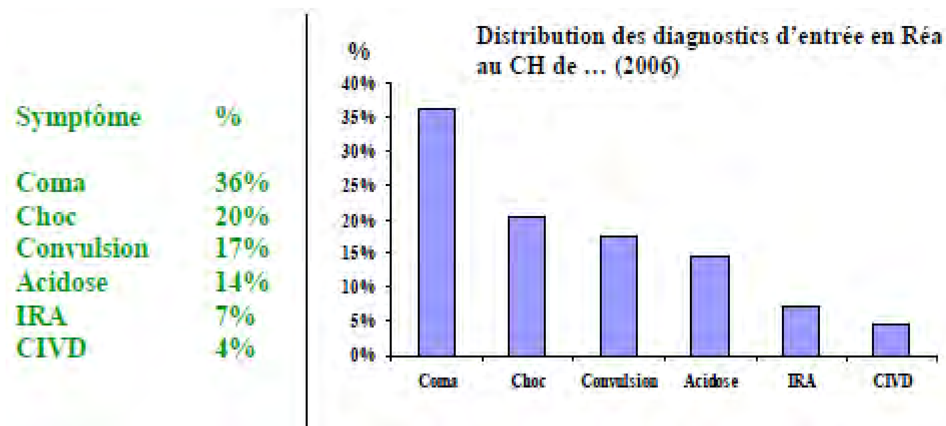
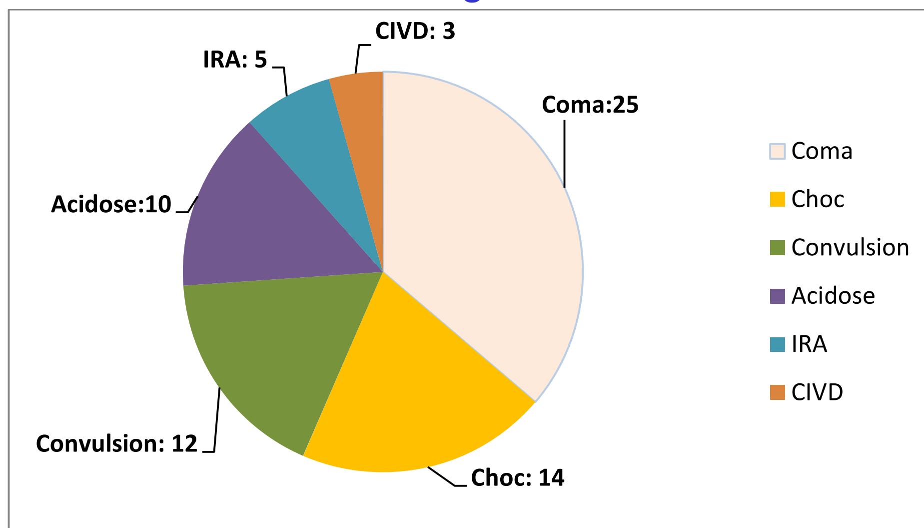
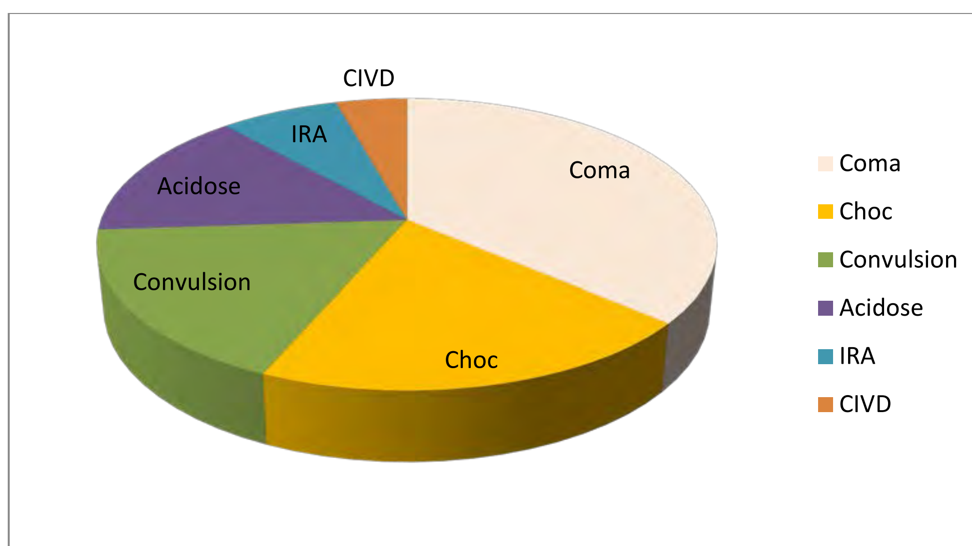


Diagramme en secteurs

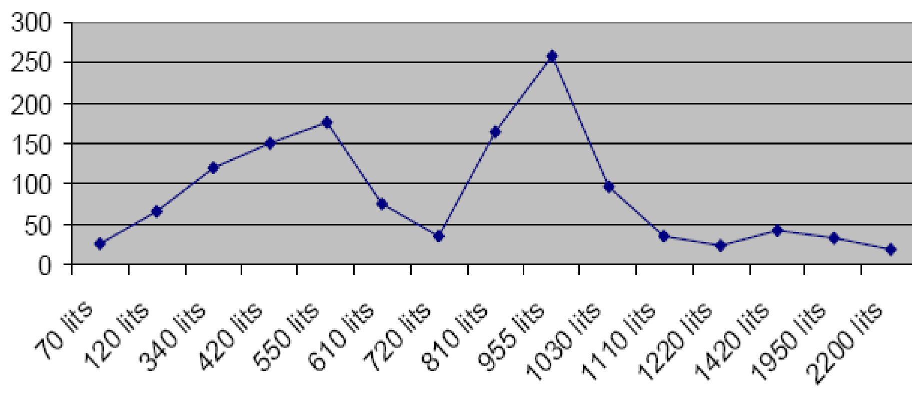




Eviter les graphes en 3 dimensions

Polygone de fréquence :

Distribution des Etablissements selon le nombre de lits



4.3.2 Présentation des données : Variables quantitative

• Distributions synthétisées par des quantités

– **de tendance centrale** (paramètres de position) : Médiane ; Moyenne ; Mode

– **de dispersion :** Variance, écart-type

4.3.2.1 Paramètres de position

4.3.2.1.1 Quartiles, déciles, percentiles

Quartiles : Les quartiles divisent la série statistique en quatre parties égales comprenant le même nombre de sujets.

Le premier quartile (quartile inférieur) est la valeur de la variable du 25ème sujet sur 100.

Le deuxième quartile n'est autre que la médiane, c'est la valeur de la variable du 50ème sujet sur 100.

Le troisième quartile (quartile supérieur) est la valeur de la variable du 75ème sujet sur 100.

Dans une série de 60 cas, par exemple, le rang du premier quartile sera 15, le rang du deuxième quartile sera 30, le rang du troisième quartile sera 45. Une fois le rang déterminé, on recherche les valeurs correspondantes de la variable.

Pour le poids de 19 étudiants, le rang du premier quartile est 4,8, le rang du deuxième quartile est 9,5, le rang du troisième quartile est 14,3.

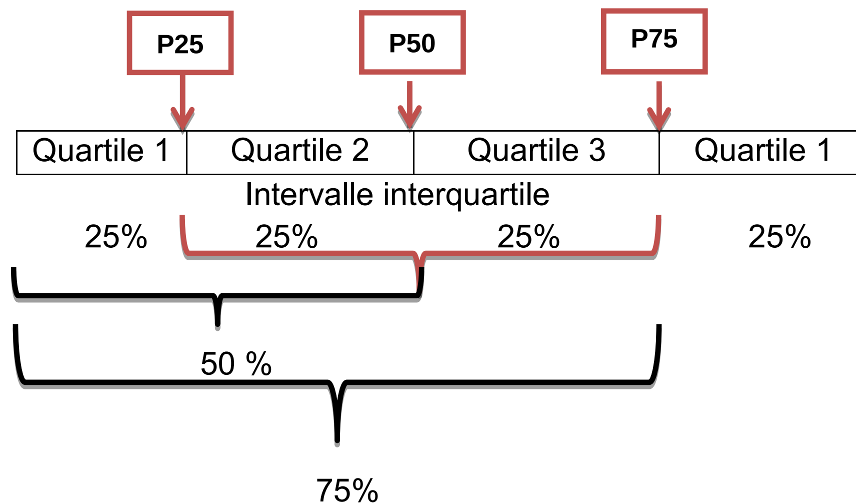
On peut à priori affecter aux différents quartiles les valeurs des centres de classes dans lesquelles ils se trouvent, soit pour notre exemple et par ordre : 62,5, 67,5, 72,5 Kg.

Il est intéressant de représenter sur graphique la distribution des fréquences relatives cumulées d'une série d'observation ceci nous permet de déterminer graphiquement la médiane.

Quartile 1 = n + 1 /4 --------- 25 % des observations < inférieures
75% des observations > supérieures

Quartile 2 = médiane

Quartile 3 = (n + 1) x 3 / 4 75 % des observations < inférieures
25% des observations > supérieures



Déciles :sont au nombre de 9. Ce sont des valeurs de la variable qui partagent la série statistique en 10 parties comprenant chacune 1/10ème de l'effectif total.
Le premier décile est la note du 10ème sujet sur 100.
Le deuxième décile est la note du 20ème sujet sur 100
Le cinquième décile se confond avec le deuxième quartile et la médiane
Le décilage est d'un usage fréquent, notamment en biométrie, parce qu'il permet de situer rapidement et facilement la position d'un sujet quelconque par rapport aux autres sujets de la série
Percentiles :sont au nombre de 99. Ce sont des valeurs de la variable qui partagent la série statistique en 100 parties comprenant chacune 1/100ème de l'effectif total.
De façon générale, les percentiles sont utilisés lorsque le nombre de valeurs de la série statistique est supérieur à 1000.
Le percentile 10% → au 1er décile
Le percentile 25 % → au 1er quartile
Le percentile 50 % à → la médiane

L'interprétation des mesures des courbes de croissance
La courbe inscrite représente des percentiles sélectionnés dansla population de référence et peut être utilisée pour repérer lerang de l'enfant par rapport aux enfants du même sexe et dumême âge.
Par exemple, lorsqu'un poids inscrit se trouve au 90^{ème}percentile du poids par rapport à l'âge, seulement 10enfants sur 100 (10 %) du même âge et du même sexe de lapopulation de référence présentent un poids plus élevé.
Les valeurs <3ème percentile ou >97ème percentile constituent desalertes (anomalies?)



4.3.2.1.2Mode :Le mode d'un ensemble de nombres est le nombre que l'on rencontre le plus fréquemment, c'est-à-dire celui qui a la plus grande fréquence ou classe de plus grande fréquence.

Le mode peut ne pas exister, et s'il existe, il peut ne pas être unique.

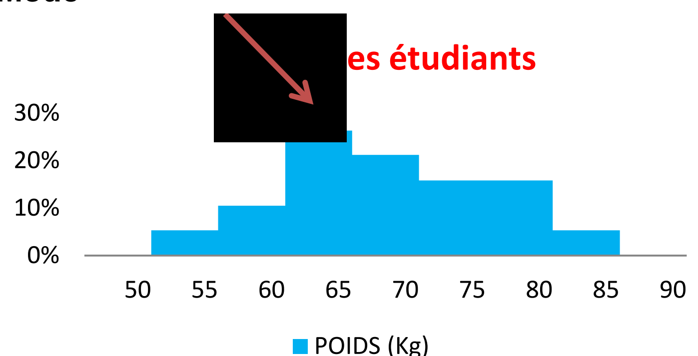
Une distribution de fréquences peut présenter un seul mode (distribution unimodale) ouplusieurs modes (distribution bi ou trimodale).

Si la distribution des valeurs est symétrique, la valeur du mode est proche de la valeur dela moyenne arithmétique. Mo ≈ x

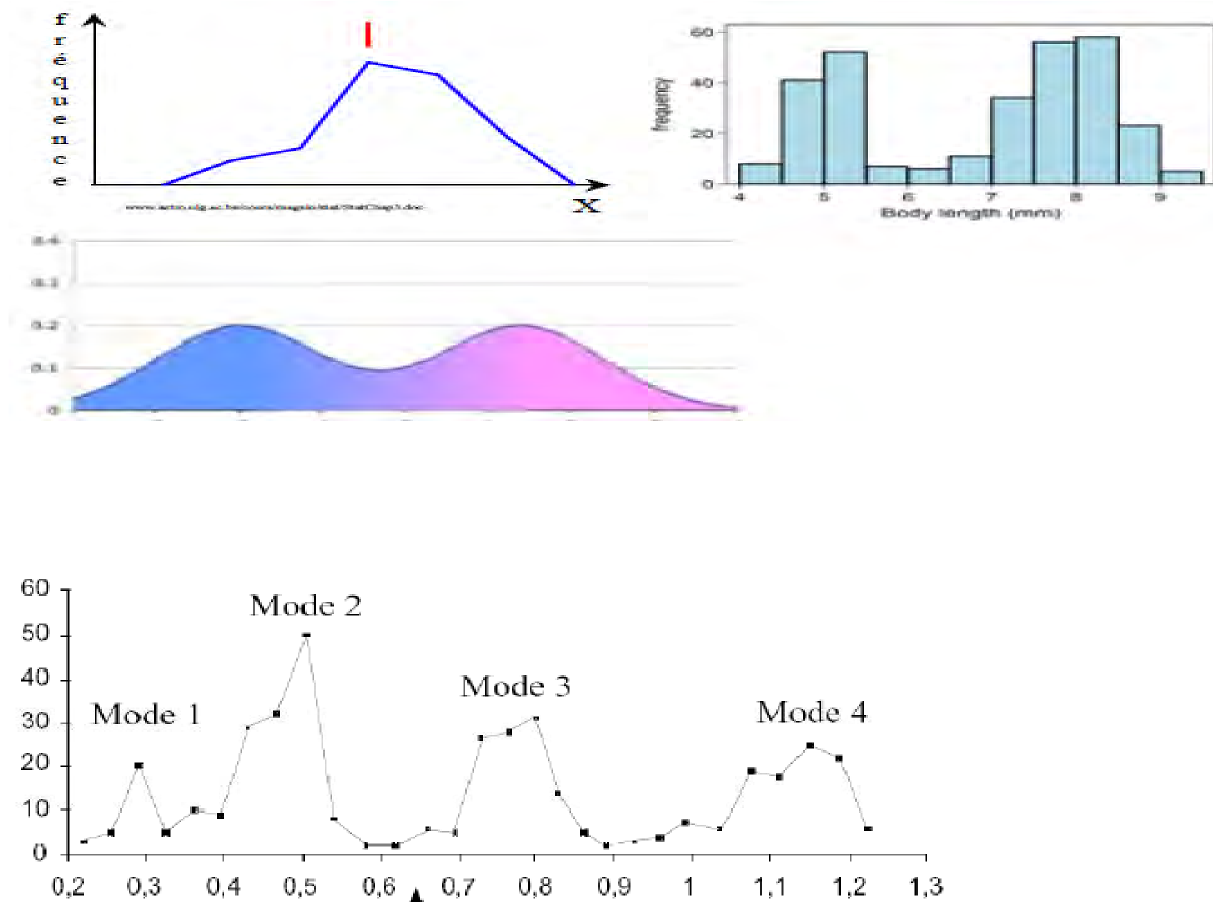
Ex: {2,10,3,3,5,3,4,1,4,2}

Valeur de	effectif		effectif cumulé
1	1		1
2	2		3
3	**3**	**Mode**	6
4	2		8
5	1		9
10	1		10

Poids (Kg)	Centre de classes	Effectif	%
50-54	52.5	1	5.3
55-59	57.5	2	10.5
60-64	62.5	**5**	26.3
65-69	67.5	4	21.2
70-74	72.5	3	15.8
75-79	77.5	3	15.8
80-84	82.5	1	5.3
Total		19	100



Distributions bimodales



Moyenne

Avantages et inconvénients

- Pas influencée par les valeurs extrêmes de la distribution des variables
- Calculable sur des caractères cycliques où la moyenne a peu de signification,
- Bon indicateur d'une population hétérogène.
- Se prête mal aux calculs statistiques,
- Très sensible aux variations d'amplitude des classes

4.3.2.1.3Médiane

Médiane = valeur de la variable qui sépare la série statistique en deux groupes d'égal effectif ou la moyenne arithmétique des valeurs centrales.

Géométriquement la médiane est lavaleur de X (l'abscisse X) correspondant à la verticale qui divise un histogramme en deuxparties d'aires égales

La médiane Me est aussi la valeur du caractère pour laquelle la fréquence cumulée est égale à50% de l'ensemble des effectifs ; elle correspond donc au centre de la série statistique classéepar ordre croissant, ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et50% sont inférieures.

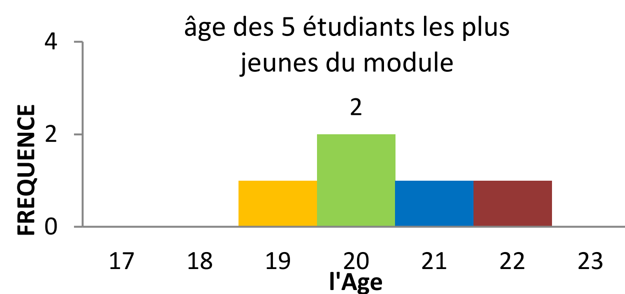
En pratique :

1. On classe les données par ordre croissant
2. La médiane est la valeur qui se trouve au milieu des données triées

Exemple : âge des 5 étudiants les plusjeunes du module (n = 5)

{19, 20, 22, 20, 21}**19, 20, 20, 21, 22**

Age	Effectif	Effectif cumulé
19	1	1
20	2	3
21	1	4
22	1	5



La Médiane = 20

Médiane = Lm +**{**(n/2 - f cum) /Fm**}** **×** i

Lm : limite inférieure de la classe dans laquelle se trouve le n/2 élément

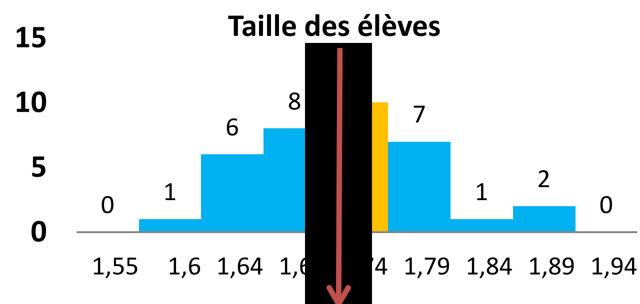
Fm: fréquence de l'élément médian

i : amplitude de classe médiane

n : effectif de l'échantillon

F cum : fréquence cumulée jusqu'à la limite inférieure de la classe médiane

Classes: taille (m)	Effectif
1.55-1.59	1
1.60-1.64	6
1.65-1.69	8
1.70-1.74	10
1.75-1.79	7
1.80-1.84	1
1.85-1.89	2
Total (n)	35



Médiane

- la série statistique suivante représente 35 élèves chez lesquels on a mesuré la tailleen m
- Valeur de la médiane= 1.70 +{ [17,5-15] /10} × 0,04 = 1,71 m

Dans le cas où les valeurs prises par le caractère étudié (variable) ne sont pas regroupées en classe :On détermine d'abord la position de la médiane.

Si N est impair :Position de la médiane : = N+1/ 2

Si N est pair : Il faut prendre la moyenne entre les deux observations

$$\frac{N}{2} \text{ et } \frac{N}{2} + 1$$

Exemple : N = 10

N représente 10 médecins chez lesquels on a mesuré la fréquence cardiaque dont voici les résultats : 58, 59, 62, 62, 65, 72, 72, 77, 77,83.

On déterminera la position de la médiane qui est placée entre les observations N/2 et N/2+1(c'est à dire entre la 5ème et la 6èmeobservation)

$$\frac{N}{2} = \frac{10}{2} = 5 \quad \text{et} \quad \frac{N}{2} + 1 = 6$$

65 + 72
-------- = 68.5 battements par minute
2

Commentaire : 50% des médecins présente une fréquence cardiaque inférieure à 68.5 battements par minute.

Avantages et inconvénients

- Pas influencée par les valeurs extrêmes de la distribution des variables
- Peu sensible aux variations d'amplitude des classes,
- Calculable sur des caractères cycliques où la moyenne a peu de signification.
- Se prête mal aux calculs statistiques,
- Suppose l'équi-répartition des données
- Ne représente que la valeur qui sépare l'échantillon en 2 parties égales.

4.3.2.1.4Moyenne : La moyenne s'exprime dans les mêmes unités que lesvaleurs observées.

Indicateur de tendance centrale servant à résumerune série de données d'une variable quantitative

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Exemple :- âge des 5 étudiants les plus jeunes du module (n = 5)
{19, 20, 20, 21, 22}

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\{19+20+20+21+22\} = \frac{102}{5} = 20.4$$

Âge des 5 premiers inscrit à l'examen
N= 5 : {19, 20, 20, 21, 42}

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\{19+20+20+21+42\} = \frac{122}{5} = \boxed{24,4}$$

N= 5 :{19, 20, 20, 21, 15}

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\{19+20+20+21+15\} = \frac{95}{5} = \boxed{19}$$

Avantages et inconvénients

- Facile à calculer,
- La somme des écarts à la moyenneest plus faible que la somme desécarts à la médiane ou au mode
- Fortement influencée par les valeurs extrêmes de la distribution des variables
- Si la distribution estdissymétrique, la moyennereprésente mal la valeur centrale.

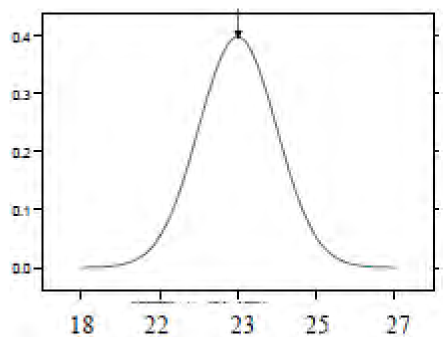
Positions relativesmédiane, moyenne, mode

• Si distribution symétriqueles : 3 coïncident

Médiane

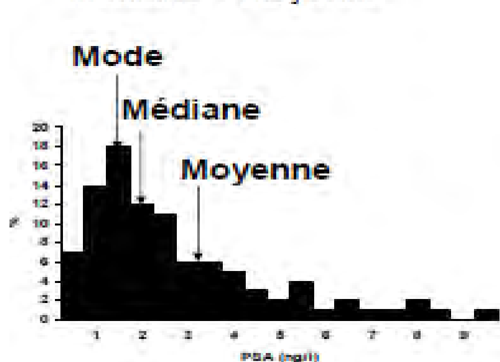
Moyenne

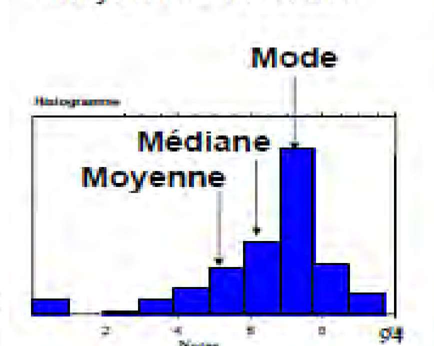
Mode



• Si distribution dissymétrique

à droite à gauche





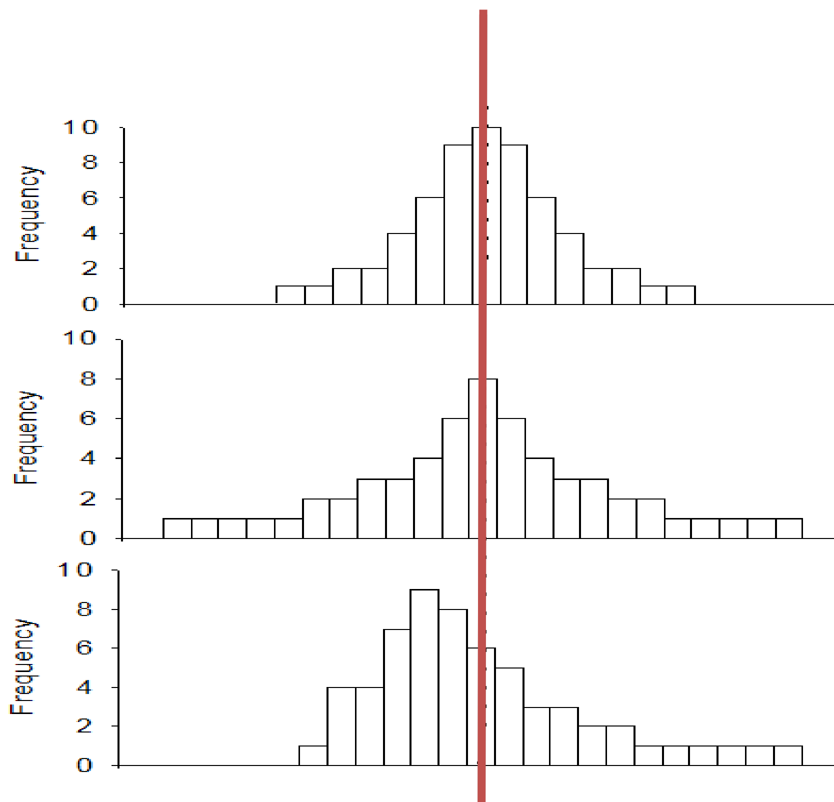
4.3.2.2 Paramètres de dispersion

La variabilité est la règle dans la les sciences de la vie

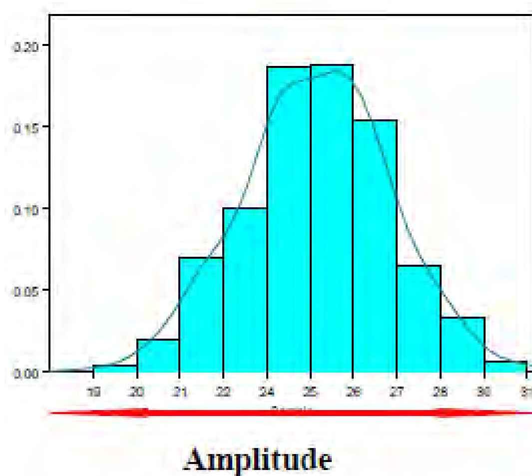
Les paramètres centraux ne résument pas complètement unedistribution.

Dispersion est la notion cléexprime la variabilité

- Les paramètres mesurant la dispersion :
- Étendue (range)
- Espace interquartile (entre 1 et 3ème)
- VARIANCE
- ECART TYPE



4.3.2.2.1 Amplitude : est la différence entre le plus grand et le plus petit de ces nombres = étendue = Valeur maximale – valeur minimale



Exemple : {2,10,3,3,5,3,4,1,4,2}

Valeur de…	effectif	effectif cumulé
1	1	1
2	2	3
3	**3 Mode**	6
4	2	8
5	1	9
10	1	10

Amplitude = 10 – 1 = 9

4.3.2.2.2 Variance :moyenne des carrés des écarts à la moyenne (écartquadratique moyen)

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\{x_i - \bar{x}\}^2 \qquad s^2 = \frac{1}{n}\left\{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}$$

Exemple : âge des 5 premiers inscrits

n = 5 {19, 20, 20, 21, 22}

$$\sum_{i=1}^{n} x_i = \{19+20+20+21+22\} = 102$$

$$\sum_{i=1}^{n} x_i^2 = \{19^2+20^2+20^2+21^2+22^2\} = 2086$$

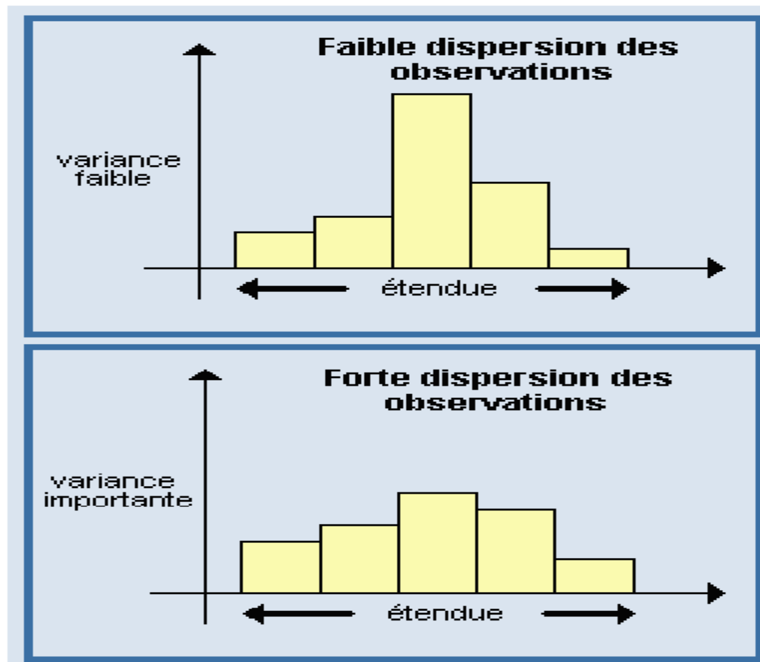
$$s^2 = \frac{1}{n}\left\{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\} = \frac{1}{5}\left\{2086 - \frac{102^2}{5}\right\} = 1.04$$

Exemple : poids d'une population de 100 femmes

Moyenne = 54.9 Kg

X	F	X- 54.9	$(X- 54.9)^2$	$F*(X- 54.9)^2$
42	5	-12.9	166.41	832.05
47	12	-7.9	62.41	748.92
52	31	-2.9	8.41	260.40
57	31	+2.1	4.41	136.71
62	16	+7.1	50.41	806.56
67	3	+12.1	146.41	439.23
72	2	+17.1	292.41	584.82
Total	100			3809

S^2 = 3809 / 100 = 38.09



4.3.2.2.3 Ecart-type : racine carrée positive de la variance

$$s=\sqrt{s^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\{x_i-\bar{x}\}^2}$$

- mesure l'écart à la moyenne
- s'exprime avec la même unité que la variable

4.3.2.2.4 Coefficient de variation

Le coefficient de variation est une mesure de dispersion desobservations d'une variable quantitative.

- C'est une mesure neutre.
- Elle est calculée en divisant l'écart-type par la moyenne.
- On exprime souvent le coefficient de variation en pourcentage.
- Sans unité, il permet de comparer facilement la dispersion desvariables différentes.

Coefficient variation= CV = Ecart type/moyenne

Exemple : 05 enfants sont mesurés par 02 observateurs A et B.

Le travail de mensuration de l'observateur A est 2 fois moins précis que celui de l'observateur B

Observateurs	moyenne	E.Type	CV
A	75	7.5	10
B	75	3.75	5

4.3.2.3Représentation graphique : variable quantitative continue

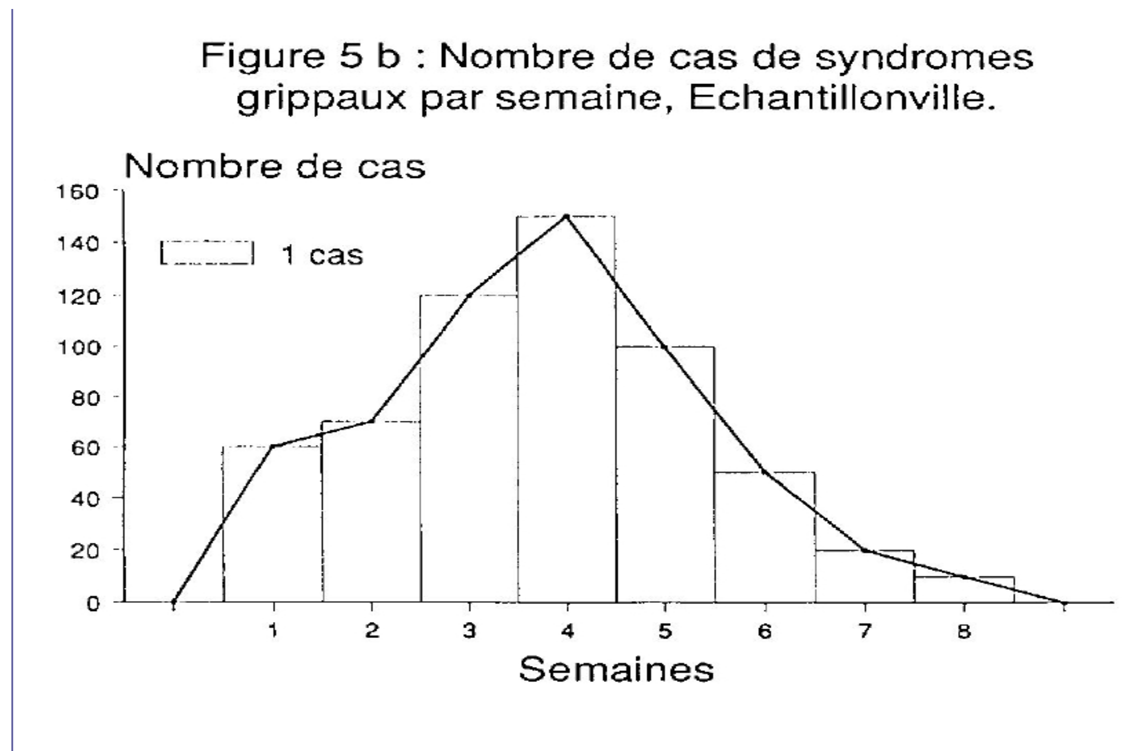
POLYGONE DE FREQUENCE :

- Construction: point de départ = histogramme

On joint simplement par une ligne le milieu des bords supérieurs des rectangles représentant chaque classe d'un histogramme

la surface sous le polygone = surface de l'histogramme

ATTENTION à la FERMETURE du polygone

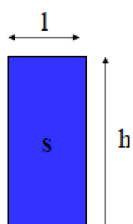


HISTOGRAMME : Histogramme présente une distribution de fréquence avec des barres verticales CONTIGÜES (ne pas confondre avec diagramme en bâton où il y a un espace entre les bâtons)

- la surface de chaque barre est proportionnelle à la valeur qu'elle représente
- surface totale = 100 % = 1

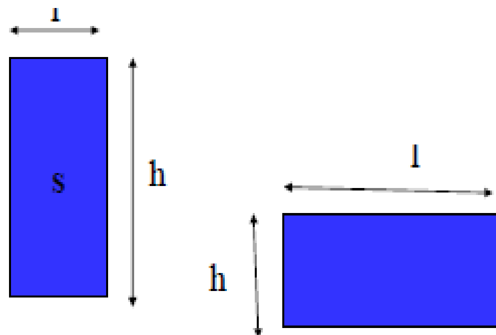
Avant toute chose on définit comment sera représenter un cas

exemple 1cm = 1surface= un cas



S = h×l (l= largeur de la classe = amplitude) = elle est proportionnelle au nombre de cas = kn ce qui compte c'est la surface = kn k pour dire c'est proportionnelle au nombre de cas

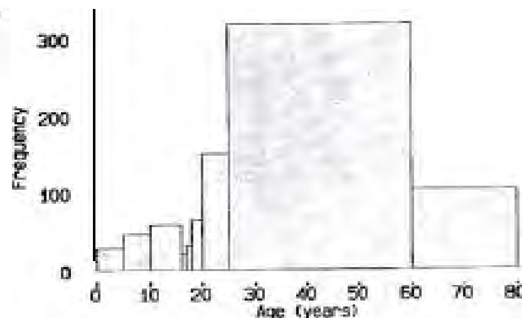
Si toutes les classes ont même largeur (l), la hauteur desrectangles est directement proportionnelle à l'effectifde la classe (h = n)



Si toutes les classes n'ont pas la même largeur (l), seulela surface des rectangles est directementproportionnelle à l'effectif de la classe

Table Road accident casualties in the London Borough of Harrow in 1985 (excluding 65 with unknown age)

	Frequency
0– 4	28
5– 9	46
10–15	58
16	20
17	31
18–19	64
20–24	149
25–59	316
60–	103
Total	[illegible]



1 cas = 1 cm^2**L'"histogramme" a**

1 an = 1 cm en abscisse

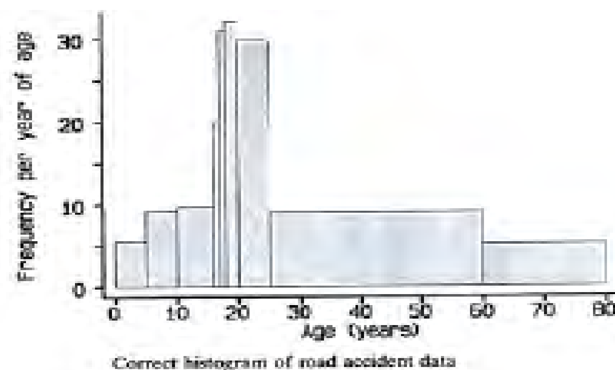
28 cas = 28 cm^2

Abscisse 0-4 =5 ans =5 cm

Hauteur = 28 : 5 = 5,6 cm

S = h x l h ⟶ S/l = n/l

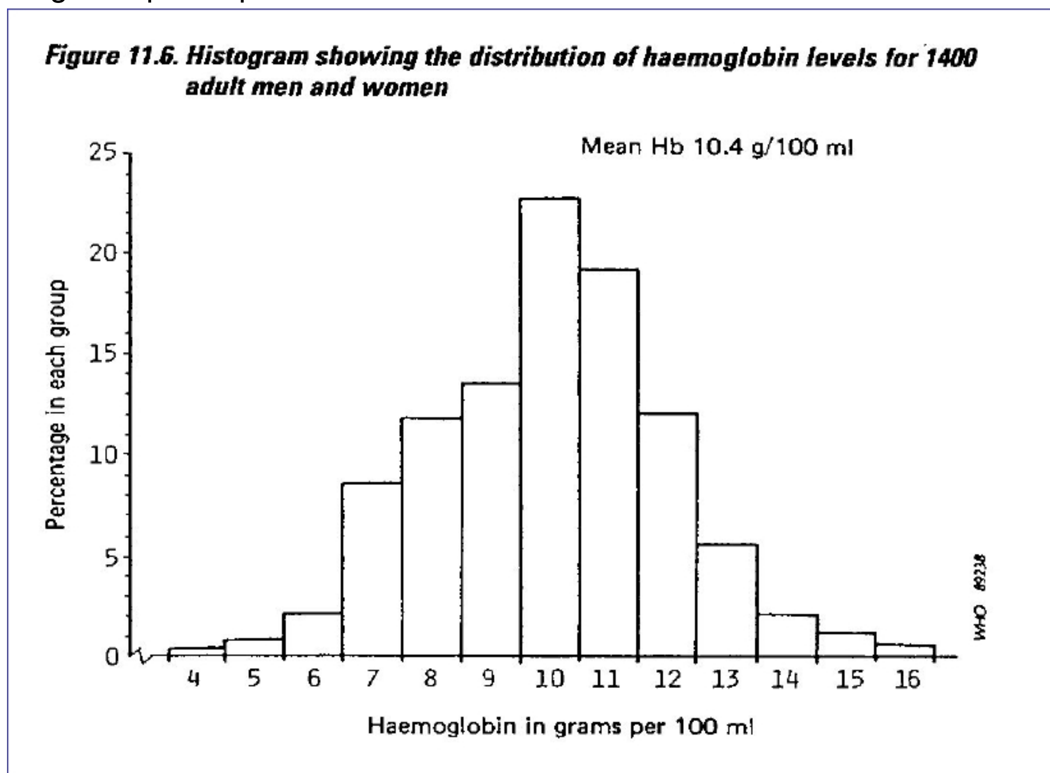
Classes	l	h
0-4	5	28/5=5.6
5-9	5	46/5=9.2
10-15	6	58/6=9.7
16	1	20
17	1	31
18-19	2	64/2=32
20-24	5	149/5=29.8
25-59	35	316/35=9.0
60-79	20	103/20=5.1



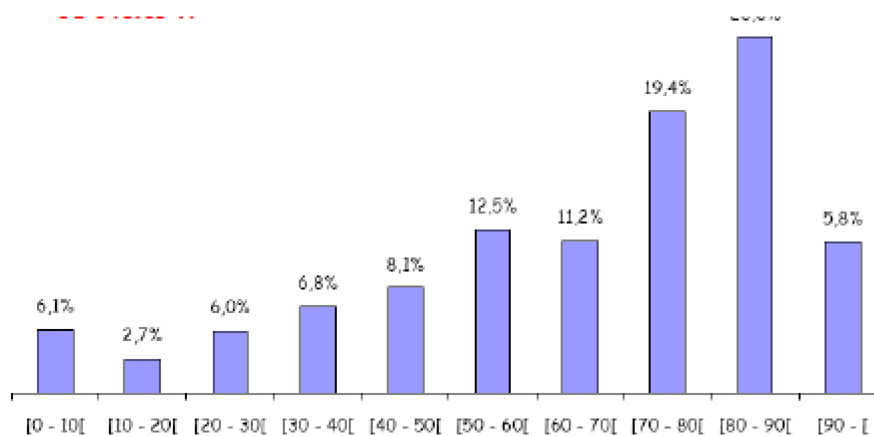
Correct histogram of road accident data

L'"histogramme" b

L'"histogramme" a (incorrect) suggère laclasse 25-59 comme étant la plus concernée parles accidents de la route alors que l'histogramme b (correct) indique que ce sont lesclasses 16, 17, 18-19. En effet, la fréquence desaccidents de la route par année d'âge est plusimportante dans ces dernières classes!



A éviter !!



Exercice 1 :

Le taux de glucose sanguin (glycémie) déterminé chez 32 sujets est donné ci-dessous en g/l :

0.85 O.87 0.90 0.93 0.94 0.94 0.95 0.97 0.97 0.98 0.98 0.99 1.00 1.01 1.03 1.03 1.03 1.04 1.06 1.07 1.08 1.08 1.10 1.10 1.11 1.13 1.14 1.14 1.15 1.17 1.19 1.20

1) De quel type de variable s'agit-il ?

2) faire une répartition en classes, en justifiant le choix du nombre de classes et l'intervalle de classe.

3) Faire la représentation graphique.

4) Calculer la moyenne et l'écart type de cette série statistique.

5) Construire la courbe des effectifs cumulés croissants.

6) Déterminer la médiane et les quartiles de cette série par la méthode graphique.

7) Déterminer le mode ou la classe modale.

Exercice 2 : les diamètres de l'induration lue sur le bras après IDR à la tuberculine chez 408 enfants d'un groupe d'allergiques spontanés au BK.

Après groupement de données en 6 classes ;

Tableau : induration lue sur le bras après IDR à la tuberculine chez 408 enfants d'un groupe d'allergiques spontanés au BK.

Limites des intervalles de classes (mm)	Centre de classe Xi	Effectifs Ni
5,5 – 10,5	8	75
10,5 – 15,5	13	118
15,5 – 20,5	18	111
20,5 – 25,5	23	55
25,5 – 30,5	28	45
30,5 – 35,5	33	4
TOTAL		N=408

1) De quel type de variable s'agit-il ?

2) Calculer la moyenne et l'écart type de cette série statistique.

3) Construire la courbe des effectifs cumulés croissants.

4) Déterminer la médiane et les quartiles de cette série par la méthode graphique.

5) Déterminer le mode ou la classe modale

Exercice 3 Une série d'observations concernant les tailles d'un groupe d'adolescents de 11 à 14 ans a donné les résultats suivants

1/ Déterminer la taille moyenne des adolescents

2/ calculer l'écart type de cette série en regroupement des données par classes

Classes Taille (xi) cm	ni
140 – 144	3
144 – 148	17
148 – 152	63
152 – 156	82
156 – 160	69
160 – 164	31
164 – 168	20
168 – 172	4
172 – 174	1
174 – 178	1
Total	